Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data

KOTIKALAPUDI SRIRAM, MEMBER, IEEE, AND WARD WHITT

Abstract—**This paper analyzes a model of a multiplexer for packetized voice and data. A major part of the analysis is devoted to characterizing the aggregate packet arrival process resulting from the superposition of separate voice streams. This is done via the index of dispersion for intervals (IDI), which describes the cumulative covariance among successive interarrival times. The IDI seems very promising as a measurement tool to characterize complex arrival processes. This paper also describes the delays experienced by voice and data packets in the multiplexer using relatively simple two-parameter approximations.**

I. Introduction

UNTIL recently, studies on communication networks for digital voice and data primarily concentrated on voice using time assigned speech interpolation (TASI), also known as digital speech interpolation (DSI), with separate channels being assigned to each voice source during each talkspurt, and data using any channels that are momentarily idle (voice has nonpreemptive priority over data) [1]-[4]. Now, packet-switched communications networks are also being developed for voice and data, with a single fast channel handling the packets from all sources [5]-[8]. Thus, significant research effort is currently being devoted to the performance of packet multiplexers for voice and data [9]-[18].

In this paper, we continue to study packet multiplexers for voice and data, giving special attention to the characteristics of the packet voice traffic. (We also consider data to some extent, but we primarily focus on voice alone.) It turns out that the aggregate packet arrival process resulting from the superposition of the streams from many voice sources is quite complicated, possessing a certain burstiness (high variability) that leads to surprisingly large packet delays in the multiplexer under heavy loads. The primary purpose of this paper is to develop a better understanding of the aggregate voice packet arrival process. We also investigate simple approximations to describe the packet delays for both voice and data.

The complexity of the aggregate voice-packet arrival process is primarily due to the bursty nature of the packet arrival process from a single voice source. The packet arrival process from a single voice source consists of arrivals occurring at fixed intervals during talkspurts and no arrivals at all during silences; see Fig. 1. Hence, the interarrival times are usually one packetization period, but occasionally much longer (one packetization period plus a silence period). As we explain in Section II-A, it is reasonable to model the packet arrival process from one voice source by a renewal process (the packet interarrival times can be regarded as i.i.d. [independent and identically distributed]), but the interarrival-time distribution in the packet arrival process from one source is highly variable, so that the renewal arrival process from each source is very bursty (in comparison to a Poisson process). As a consequence, the aggregate packet arrival process resulting from the superposition of many independent voice packet streams is not nearly a renewal process. As others have observed before, this is to be expected because the instantaneous arrival rate in the aggregate packet voice arrival process at any time is a function of the number of voice sources in talkspurts, which fluctuates substantially.

The main idea here is to focus on the dependence among successive interarrival times in the aggregate packet arrival process. In particular, we apply the index of dispersion for intervals (IDI) [19, pp. 71-72]. This technique applies much more broadly, and we believe that it can greatly help understand other complex arrival processes in queueing systems (and elsewhere). Let $\{X_k, k \geq 1\}$ represent the sequence of interarrival times of an arrival process. We assume that $\{X_k, k \geq 1\}$ is stationary, by which we mean that the joint distribution of $(X_k, X_{k+1}, \cdots, X_{k+m})$ is independent of k for all m. Let $S_k = X_1 + \cdots + X_k$ denote the sum of k consecutive interarrival times. The IDI, which we also call the k-interval squared coefficient of variation sequence, is the sequence $\{c_k^2, k \geq 1\}$ defined by

$$c_k^2 = \frac{k \operatorname{Var}(S_k)}{[E(S_k)]^2} = \frac{\operatorname{Var}(S_k)}{k[E(X_1)]^2} = c_1^2 + \frac{\sum\limits_{\substack{i,j=1 \\ i \neq j}}^{k} \operatorname{cov}(X_i, X_j)}{k[E(X_1)]^2}$$

$$= \frac{k \operatorname{cov}(X_1, X_1) + 2 \sum\limits_{j=1}^{k-1} (k-j) \operatorname{cov}(X_1, X_{1+j})}{k[E(X_1)]^2},$$

$$k \geq 1, \tag{1}$$

Manuscript received November 5, 1985; revised March 19, 1986. This paper was presented in part at the IEEE Global Telecommunications Conference, New Orleans, LA, December 1985.

The authors are with AT&T Bell Laboratories, Holmdel, NJ 07733.

IEEE Log Number 8609926.

0733-8716/86/0900-0833$01.00 © 1986 IEEE

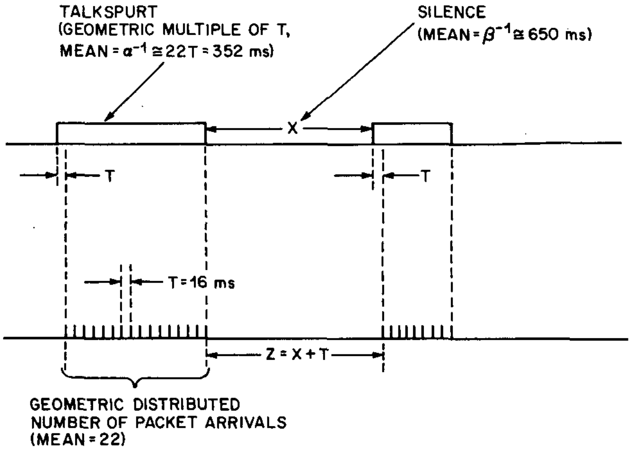

TALKSPURT
(GEOMETRIC MULTIPLE OF T,
MEAN = $\alpha^{-1} \cong 22T = 352$ ms)
SILENCE
(MEAN = $\beta^{-1} \cong 650$ ms)
X
T
T
T = 16 ms
Z = X + T
GEOMETRIC DISTRIBUTED
NUMBER OF PACKET ARRIVALS
(MEAN = 22)

Fig. 1. Packet arrival process due to one voice source.

where cov (X_i, X_j) is the covariance between X_i and X_j, defined by cov $(X_i, X_j) = E(X_iX_j) - E(X_i)\,E(X_j)$. For $k = 1$, $c_k^2 = c_1^2$ is the squared coefficient of variation (variance divided by the square of the mean) of a single interarrival time. For $k > 1$, c_k^2 is k times the squared coefficient of variation of S_k. The factor k appears because the squared coefficient of variation of S_k typically converges to 0 as $k \to \infty$, due to the law of large numbers. For $k > 1$, c_k^2 measures the cumulative covariance (normalized by the square of the mean) among k consecutive interarrival times. The notion of cumulative covariance seems to be very important for the multiplexer application, because the exceptionally large packet delays under heavy loads are evidently due, not to high values of c_1^2 or cov (X_i, X_j), but to the cumulative effect of many small individual covariances.

We were motivated to use the IDI because its limiting value $c_\infty^2 = \lim_{k\to\infty} c_k^2$ is known to completely characterize the effect of a general stationary arrival process (beyond its average arrival rate) on the congestion of a multiserver FIFO queue in heavy traffic; see [20, Theorem 1] and [21, Theorem 20.1]. (The limiting term c_∞^2 is essentially, except for the choice of measuring units, the normalization constant that appears in the central limit theorem for $\{S_k, k \geq 1\}$.) The IDI (actually Var (S_k)) is suggested as a tool to develop approximations for arrival processes to queues in [22, Section 2]. The covariances were also used for a related problem by Heffes [23] and have been applied by Heffes and Lucantoni [24] to develop a different approximation for this model. In their work, the superposition arrival process is approximated by a Markov-modulated Poisson process (MMPP), and the resulting queueing model is solved using matrix-analytic methods [25].

The specific model we consider for the multiplexer is a single-server queue with unlimited waiting room and the first-in-first-out (FIFO) service discipline. The inputs to the multiplexer consist of aggregated voice traffic streams together with data traffic. We assume that the service times are independent of the arrival process and i.i.d. with a general distribution. (The service-time distribution is often deterministic.) We also assume that the overall packet arrival process can be represented as a superposition of independent renewal processes, so that in standard queueing parlance we have a $\Sigma GI_i/G/1$ model. The most questionable assumption for applicability is the service discipline, because many voice/data multiplexers use priority schemes. For simplicity, we consider only a FIFO model here, but it seems clear that the insights about voice packet traffic generalize. (For example, the limiting quantity c_∞^2 in the IDI is known to play a similar role for priority queues in heavy traffic [26], [27].)

With the modeling assumptions above, analyzing the performance of the multiplexer reduces to analyzing a $\Sigma GI_i/G/1$ queueing model, but the superposition arrival process makes this an extremely difficult model to analyze exactly. Thus, we resort to approximations. In particular, we use two-parameter approximations as in the queueing network analyzer (QNA) software package [22], [28]–[32]. With this technique, the superposition arrival process is approximately characterized by two parameters, one representing the average arrival rate and the other the variability. The variability parameter can be viewed as the squared coefficient of variation (variance divided by the square of the mean) of an interarrival time in an approximating renewal arrival process, but this does not mean that the dependence in the original process is ignored. The procedure attempts to capture the dependence (the covariances) by making the variability parameter depend upon elementary properties of the IDI. Since the relevant covariances tend to depend on the traffic intensity in the queue, the approximating variability parameter also depends on the traffic intensity of the queue, even though the arrival process is an exogenous input (independent of the service times in the queue). The primary advantage of this approach is that it readily produces approximations that capture the main qualitative behavior. The closed-form formulas also help to provide insight.

This approximation approach was previously applied by Jenq [15] to analyze a multiplexer serving only packet voice. Our work began with the intent of extending Jenq's analysis to include data as well. Since the QNA approximation formulas in [28] apply to arbitrary superposition arrival processes, this extension seemed to present no difficulties, but an anomaly in the QNA formulas was discovered when the data is modeled as a Poisson process. Since the superposition of independent Poisson processes is again Poisson, it should not matter whether we represent the Poisson data as the superposition of n independent Poisson data streams each with rate λ or a single Poisson data stream with rate $n\lambda$. Unfortunately, however, the QNA formulas are not independent of this specification when there are also non-Poisson voice processes present. (See Section III-A). We found that the approximation tends to perform best if the number n of Poisson data streams is selected so that the rate λ of each individual Poisson data stream is approximately the same as the average arrival rate of each individual non-Poisson voice arrival stream. Upon reflection, this is intuitively reasonable; we should expect that the QNA approximation for superposition processes would perform better if the processes being superposed are all approximately in the same time scale (have similar rates).

It soon became clear that the notion of *relevant time scale* is very important more generally. From the point of view of classical superposition limit theorems [33] (discussed in Section II), we should expect that the superposition of many component stationary arrival processes would be nearly Poisson, but even with more than a hundred voice sources, the multiplexer under heavy loads experiences packet delays much greater than would occur with a Poisson arrival process. In fact, the deviation from Poisson behavior increases as the number of streams increases. This apparent contradiction can be explained by focusing on the time scale. The notion of time scale appears indirectly in the classical superposition limit theorem [33] in the requirement that the individual streams get sparse as the number of streams increases, so that the total average arrival rate remains unchanged. This condition is clearly violated in the multiplexer example, but we could rescale time in the aggregate arrival process to enforce this condition. The real issue then is the relevant time scale for the aggregate arrival process in its intended application. We will show that the aggregate voice packet arrival process does behave like a Poisson process over relatively short time intervals, but under heavy loads the congestion in the multiplexer is determined by the behavior of the arrival process over much longer time intervals, where it does not behave like a Poisson process. Much of this paper is devoted to clarifying this phenomenon of time scale for arrival processes to queues. Related discussion appears in [34].

This paper extends Jenq [15] not only by considering multiplexers for data as well as voice, but also by describing the probability of delay and the standard deviation of delay as well as the expected delay. ("Delay" or "waiting time" here refers to the time required to reach the server; it does not include service time.) We also further investigate the quality of the QNA approximations, using the IDI for guidance. In particular, in Section III-C-2), we develop a new modification of the QNA approximation especially for the multiplexer model. Thus the paper has two principal thrusts: first, to characterize the aggregate packet arrival process using the IDI and, second, to extend and further evaluate the simple QNA approximations.

The rest of this paper is organized as follows. In Section II we focus on the arrival process. Section II-A describes our renewal-process model for the packet-arrival process from a single voice source; Section II-B discusses indexes of dispersion and characterizes the superposition arrival process associated with multiple voice sources; and Section II-C discusses the interaction between the arrival process and the queue (the issue of time scale). In Section III, we present and evaluate approximations describing packet delays in the multiplexer. Section III-A reviews some of the QNA formulas in [28]; Section III-B describes the simulation experiments and makes numerical comparisons; and Section III-C presents the new improved approximations for this model. In Section IV, we consider the related model with a finite buffer, to test the hypothesis that the arrival process behaves like a Poisson process for small buffer sizes but not for large buffer sizes, because of the time scale. The idea is that a small buffer should greatly restrict the interactions among arrivals widely separated in time, reducing the impact of many small long-term covariances. Finally, in Section V, we present our conclusions, emphasizing the usefulness of the IDI to describe the variability of arrival processes.

II. Characterization of Packetized Voice Traffic

A. Single Voice Source

The packet stream from a single voice source is characterized by arrivals at fixed intervals of T ms during talkspurts and no arrivals during silences (see Fig. 1). We assume that successive talkspurts and silence periods form an alternating renewal process; i.e., all these time intervals are independent with each talkspurt being of random length NT and each silence period of random length X. Our most important assumption is that the number N of packets in a talkspurt is geometrically distributed on the positive integers. This is consistent with measurements indicating that talkspurts are approximately exponentially distributed [35]-[38]. Because of the lack of memory property associated with the geometric distribution, the packet stream due to a single voice source is a renewal process. To specify our model completely, we assume that the voice packetization period is $T = 16$ ms, the silence periods are exponentially distributed with mean $\beta^{-1} = 650$ ms, and that the mean number of packets per talkspurts is $E(N) = 22$, so that the mean talkspurt is $\alpha^{-1} = 352$ ms [38]. (An exponential distribution for the silence period seems reasonable given the measurement studies, although it is not a perfect fit [35]. In fact, any distribution for X could be used in our analysis.) In other words, each packet interarrival time from one voice source is of length $T = 16$ ms with probability $p = \frac{21}{22}$ and of length $X + T$ (where $E(X + T) = 666$ ms) with probability $1 - p = \frac{1}{22}$, as shown in Fig. 2. (The interarrival time including a silence is $X + T$ because the first packet in a talkspurt does not arrive until after T ms. The voice-packet size depends on the coding scheme used; e.g., for 32 kbit/s ADPCM coding and $T = 16$ ms, the packet size is 64 bytes.)

The squared coefficient of variation (variance divided by the square of the mean) of an interarrival time in this renewal process is

$$c_1^2 = (1 - p^2)/[T\beta + (1 - p)]^2 = 18.1. \qquad (2)$$

Even though the packet arrival stream from a single voice source can be modeled by a renewal process, it is very bursty, as is reflected by the very high value of c_1^2.

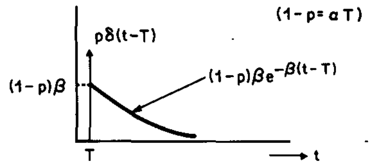


Fig. 2. Probability density function of packet interarrival time for one voice source.

B. Multiple Voice Sources

We now proceed to characterize the aggregate packet arrival process resulting from the superposition of n packet-voice sources in terms of the indexes of dispersion for intervals and counts; [19, pp. 71-72]. The index of dispersion for intervals (IDI) was defined in (1); now we define the index of dispersion for counts (IDC).

Let $N(t)$ denote the counting process associated with an arrival process. (See [19] or [22, Section 1] for additional background.) The index of dispersion for counts, $I(t)$, is the function

$$I(t) = \frac{\text{Var}\,[N(t)]}{E[N(t)]}, \; t > 0. \tag{3}$$

The following proposition summarizes well known properties of the two indexes of dispersion.

Proposition 1): a) *For a Poisson process,* $I(t) = 1 = c_k^2$ *for all t and k.* b) *For a renewal process,* $c_k^2 = c_1^2$ *for all k.* c) *If* $c_k^2 = c_1^2$ *for all k, then* cov $(X_i, X_j) = 0$ *for all* $i, j \; (i \neq j)$.

Looking for fluctuations in the IDI sequence $\{c_k^2, k \geq 1\}$ is a good way to test for deviations from the renewal property. In [22], [28], [29], the sequence $\{c_k^2\}$ is used as the basis for calculating the variability parameter to approximately characterize the arrival process. The stationary-interval method in [22] uses c_1^2; the asymptotic method in [22] uses c_∞^2; and Albin's hybrid methods in [29] use the convex combinations $wc_\infty^2 + (1 - w)c_1^2$. More generally, it is natural to use weighted functions of the entire sequence, i.e., $\Sigma_{k=1}^{\infty} w_k c_k^2$ where $w_k \geq 0$ and $\Sigma_{k=1}^{\infty} w_k = 1$, but such procedures still need to be developed.

Let c_{kn}^2 and $I(t; n)$ denote the two indexes of dispersion associated with the superposition process of n independent and identically distributed arrival processes. Important insight into the superposition process is obtained from two limits, one as $t \to \infty$ or $k \to \infty$ with fixed n, and the other as $n \to \infty$ with fixed t or k; see [19] and [39, pp. 221-229].

Proposition 2): For a superposition of n independent and identically distributed renewal processes,

$$I(\infty; n) \equiv \lim_{t \to \infty} I(t; n) = c_{\infty n}^2 = \lim_{k \to \infty} c_{kn}^2 = c_{11}^2 \tag{4}$$

for any fixed n, where c_{11}^2 *is the squared coefficient of variation of a single interval in one of the renewal processes being superposed.*

Proposition 3): The superposition of n independent and identically distributed renewal processes each with rate λ/n *tends to a Poisson process with rate* λ *as n tends to infinity.*

In fact, Proposition 3) also applies to nonrenewal component processes; see Cinlar [33]. A key condition in Proposition 3) is that the processes being superposed become increasingly sparse as n increases. However, we are considering a superposition of packet-voice streams with fixed individual average arrival rates, independent of n. In fact, the IDC in the superposition of n packet-voice sources with fixed individual arrival rates is identical to that of a single source:

$$I(t; n) = \frac{\text{Var}\left[\sum_{i=1}^{n} N_i(t)\right]}{nE[N_1(t)]} = \frac{\text{Var}\,[N_1(t)]}{E[N_1(t)]} = I(t; 1), \tag{5}$$

where $N_i(t)$ is the counting process corresponding to the ith source. Equation (5) suggests that the superposition of packet-voice arrival processes can never be regarded as a Poisson process. However, Proposition 3 can be invoked to deduce that the joint distribution of any fixed number of interarrival times in the superposition of packet-voice processes tends to the distribution of independent exponential variables (a Poisson process) as n increases. The apparent contradiction is removed by observing that the expected interarrival time in the superposition of n processes is $(EX_1)/n$, where EX_1 is the expected value for one stream. Focusing on a time scale in which the expected interarrival time is fixed as n changes is equivalent to making the individual streams sparse in Proposition 3); i.e., we are then considering $I(t/n; n)$.

In fact, if we do not consider the random-environment view (i.e., that there is a randomly varying instantaneous arrival rate due to a varying number of sources in simultaneous talkspurt), familiarity with Proposition 3) might well lead one to expect that a Poisson approximation for the superposition process ought to perform quite well, because the number of component streams here is quite large, about 100. In fact, contrary to what one might expect from the random-environment view alone, the Poisson approximation *does* work well under light-to-moderate loads. However, a Poisson approximation for the arrival process seriously underestimates delays under higher loads, where the long-term covariances matter. For a comparison, see Fig. 6 and Table III. (The numerical example is described in Section III-B).

The extent to which a superposition process is nearly Poisson thus depends not only on n, but also on the relevant time scale. (See [34, pp. 536, 543] for further discussion.) In our case, over short intervals of time the superposition process looks like a Poisson process; in fact, over short intervals of time the superposition process looks like something slightly less variable than a Poisson process; but over longer intervals of time the superposition process significantly deviates from a Poisson process and is highly variable. This can be seen from the IDC, $I(t; n)$, which has been calculated analytically for this superpo-

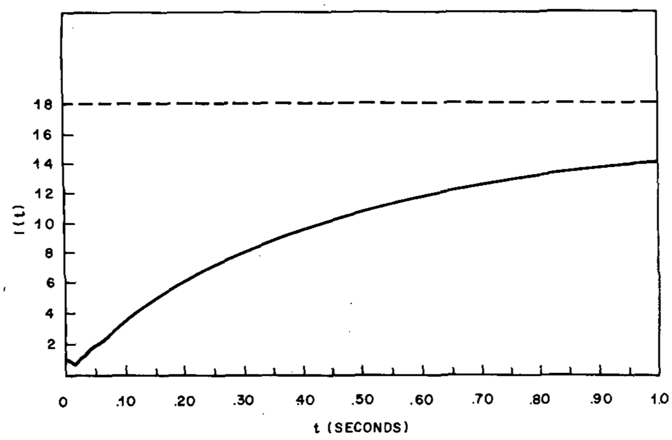


Fig. 3. Index of dispersion for number of arrivals in (0, *t*) for voice-packet arrival process.

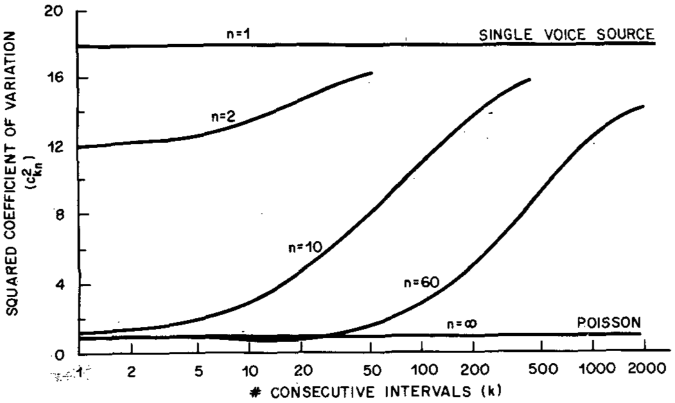


Fig. 4. k-interval squared coefficient of variation curves for superposition of n voice sources.

TABLE I

THE TAIL PROBABILITIES OF THE STATIONARY-INTERVAL DISTRIBUTION IN THE VOICE SUPERPOSITION ARRIVAL PROCESS: A COMPARISON OF THE EXACT THEORETICAL VALUES TO AN EXPONENTIAL DISTRIBUTION HAVING THE SAME MEAN. (THERE ARE n SOURCES, EACH WITH AVERAGE ARRIVAL RATE = 0.0223 PACKETS/ms.)

	n = 20		n = 100	
TIME (ms) t	STATIONARY INTERVAL $1-F_n(t)$	EXPONENTIAL $e^{-n\lambda t}$	STATIONARY INTERVAL $1-F_n(t)$	EXPONENTIAL $e^{-n\lambda t}$
0.1	0.9591	0.9570	0.8044	0.8028
0.2	0.9198	0.9159	0.6468	0.6446
0.4	0.8457	0.8389	0.4175	0.4155
1.0	0.6558	0.6446	0.1110	0.1113
2.0	0.4260	0.4155	0.0117	0.0124
4.0	0.1743	0.1726	0.00011	0.00015
16.0	0.00027	0.00088	2.5×10^{-19}	5.5×10^{-16}

sition process (for the specific data in Section III-B with $c^2_{11} = 18.1$) by Heffes and Lucantoni [24]; see Fig. 3.

A study of the k-interval squared coefficient of variation (the IDI) as a function of k and n consolidates the observations made above. The c^2_{kn} versus k values are shown in Fig. 4 for the packet-voice superposition process for various values of n. These results were obtained by extensive simulations on a Cray-1. (See Section III-B-1 for further discussion about the simulation experiment.) From Fig. 4, it is apparent that $c^2_{kn} \rightarrow 1$ as $n \rightarrow \infty$ for all k (not monotonically). Fig. 4 is also consistent with Proposition 2), showing that $c^2_{kn} \rightarrow c^2_{11}$ as $k \rightarrow \infty$ for all n. Thus, it is observed that as n increases, increasingly more consecutive intervals in the superposition process are nearly uncorrelated. However, for a fixed n, interval covariances are quite significant and positive valued when a sufficiently large number of consecutive intervals are viewed collectively. These covariances account for the rising portion of the curves in Fig. 4.

The fact that the IDI sequence $\{c^2_{kn}, k \geq 1\}$ is not nearly constant in k for large n clearly demonstrates that the superposition process is not nearly a renewal process (Proposition 1-b). The superposition operation has changed the location of the variability. The exceptional variability (deviation from a Poisson process) in each component stream (from one voice source) is entirely in the interarrival-time distribution; the interarrival times are independent. In contrast, the exceptional variability in the superposition process is almost entirely due to the long-term covariances. Nearby interarrival times are nearly independent, which is indicated by the flat left-hand portion of the IDI curves in Fig. 4. Further, the interarrival-time distribution in the superposition process is nearly exponential. We substantiate this by calculating the exact theoretical interarrival-time distribution in the aggregate packet arrival process and its squared coefficient of variation c^2_{1n} using well-known formulas, e.g., [22, eq. (4.4) and (4.5)]. (Supporting details are given in the Appendix.) The resulting formula is (see Section II-A for definitions of the parameters)

$$c^2_{1n} = 1 - \frac{2}{n+1} + \left(\frac{1-p}{T\beta + 1 - p}\right)^{n+1} \left(\frac{2}{1-p} - \frac{2n}{n+1}\right). \tag{6}$$

The exact theoretical interarrival-time distribution in the aggregate packet arrival process is compared to an exponential distribution with the same mean for $n = 20$ and 100 in Table I; c^2_{1n} as a function of n is depicted in Table II. Table II shows that the last term in (6) becomes negligible for large n. Table II also shows that c^2_{1n} decreases rapidly from 18.1 for $n = 1$ to about 1 at $n = 10$, then continues below 1, reaching a minimum of 0.906 for $n = 18$, and then approaches 1 as $n \rightarrow \infty$, consistent with Proposition 3).

TABLE II
THE SQUARED COEFFICIENT OF VARIATION OF THE STATIONARY-INTERVAL DISTRIBUTION IN THE VOICE SUPERPOSITION ARRIVAL PROCESS: THE EXACT THEORETICAL VALUE AS A FUNCTION OF THE NUMBER n OF SOURCES. (THE MINIMUM VALUE OF c_{1n}^2 IS MARKED WITH AN ASTERISK.)

n	$\frac{n-1}{n+1}$	c_{1n}^2
1	0.000	18.10
2	0.333	11.98
4	0.600	5.47
8	0.778	1.64
10	0.818	1.18
12	0.846	0.998
14	0.867	0.930
16	0.882	0.909
18	0.895	0.906*
20	0.905	0.908
22	0.913	0.914
32	0.939	0.939
64	0.969	0.969
128	0.9845	0.985
256	0.9922	0.993

These observations have important implications for statistical measurements. Note that the aggregate packet arrival process for large n will pass many tests for the Poisson process with flying colors. It is also important to look for the cumulative effect of small long-term covariances.

C. Interaction With the Queue

The above results suggest that the superposition arrival process, for fairly large n, will affect the queue like a Poisson process at lower traffic intensities, but more like a single highly variable component renewal process under higher traffic intensities, because the long-term covariances among interarrival times begin to affect the queue under higher traffic intensities. In fact, there are exact analytical results regarding queues with general stationary arrival processes at high traffic intensities that provide theoretical support for this description in heavy traffic [20], [31], [32]. At high traffic intensities, the covariances over many interarrival times significantly influence the queue behavior. In fact, in the heavy-traffic limit the impact of the arrival process on the queue is determined by c_∞^2. This is justified for a stationary sequence of interarrival times by combining [20, Theorem 1] with [21, Theorem 20.1].

In fact, we are primarily interested in the performance of the multiplexer as a function of the number n of voice lines. From Proposition 3) alone, we might expect the Poisson approximation to get better as n increases, but it does not; it gets much worse (Fig. 6 and Table III). We have seen that this can be explained partly by the fact that the individual streams are not getting sparse an n increases, but they can be regarded as getting sparse if we adjust (rescale) the time scale with n. From the point of view of the queue, what matters is how the traffic intensity ρ changes with n as n increases. Since the mean service time is fixed, here ρ is directly proportional to n. The heavy-traffic theory in [31] and [32] indicates that $n(1-\rho)^2$ is critical as both $n \to \infty$ and $\rho \to 1$. Here $n(1-\rho)^2 \to 0$ as $n \to \infty$, so that the limit as $n \to \infty$ is essentially the same as $\rho \to 1$ with n fixed. In other words, here the limit in Proposition 2) eventually dominates the limit in Proposition 3) as $n \to \infty$. As n increases and $\rho \to 1$, the traffic interaction in the queue spans over many intervals in the superposition arrival process. Thus, the long-term covariances between the interarrival times in the aggregate packet arrival process (as indicated by the rising portion of the IDI curves in Fig. 4) play a significant role, and the aggregate packet arrival process eventually looks substantially more variable than a Poisson process.

III. PERFORMANCE ANALYSIS OF THE MULTIPLEXER

A. The Queueing Network Analyzer (QNA) Technique

The multiplexer is modeled as a standard single-server queue with unlimited waiting room and the FIFO queueing discipline (i.e., no priority is given to either voice or data). The traffic entering the FIFO queue is a superposition of voice and data packet streams. Assuming that the outgoing transmission rate is R kbit/s, a packet service time is X/R ms when X is the packet size in bits. With our analysis techniques, the packet size X can be a random variable with a general distribution. If there is no data traffic, then it is natural to let the service time be deterministic because voice packets are typically all the same size.

Ten parameters are used to represent the voice and data lines: n_1, n_2, λ_1, λ_2, c_1^2, c_2^2, τ_1, τ_2, c_{s1}^2 and c_{s2}^2. There are n_1 (n_2) independent and identically distributed voice (data) lines. The arrival rate of each voice (data) line is λ_1 (λ_2) and the variability parameter or squared coefficient of variation of the interarrival times in each voice (data) line is c_1^2 (c_2^2). The total arrival rate is thus $\lambda = n_1\lambda_1 + n_2\lambda_2$. The mean and squared coefficient of variations of the packet length for voice (data) are τ_1 and c_{s1}^2 (τ_2 and c_{s2}^2). Let τ and c_s^2 be the mean and the squared coefficient of variation associated with the packet service time for the combined voice and data traffic, computed by

$$\tau = \frac{n_1\lambda_1\tau_1 + n_2\lambda_2\tau_2}{n_1\lambda_1 + n_2\lambda_2}$$

and

$$(c_s^2 + 1)\tau^2 = \frac{n_1\lambda_1(c_{s1}^2 + 1)\,\tau_1^2 + n_2\lambda_2(c_{s2}^2 + 1)\tau_2^2}{(n_1\lambda_1 + n_2\lambda_2)}, \quad (7)$$

using standard formulas for mixtures; i.e., the kth moment of a mixture of distributions is a mixture of the kth moments. The transmission line utilization is thus $\rho = \tau\lambda = n_1\lambda_1\tau_1 + n_2\lambda_2\tau_2 = \rho_1 + \rho_2$.

The QNA approximation [28] proceeds in two steps. First, the complicated superposition arrival process is approximated by a renewal arrival process partially characterized by the first two moments of its interarrival-time distribution or, equivalently, by the mean λ^{-1} and squared coefficient of variation c_a^2 of the interarrival-time distribution.

Second, approximation formulas are applied for the various congestion measures in a *GI/G/*1 queue partially characterized by the first two moments of the interarrival-time and service-time distributions, i.e., the parameter four-tuple $(\lambda, c_a^2, \tau, c_s^2)$. The following is the formula for the squared coefficient of variation, c_a^2, of the interarrival-time distribution in the approximating renewal process for the aggregate packet arrival process entering the multiplexer queue:

$$c_a^2 = wc_{\mathrm{AM}}^2 + (1 - w), \quad c_{\mathrm{AM}}^2 = \frac{n_1\lambda_1 c_1^2 + n_2\lambda_2 c_2^2}{n_1\lambda_1 + n_2\lambda_2},$$

$$w(\rho, v) \equiv w = [1 + 4(1 - \rho)^2(v - 1)]^{-1}$$

and

$$v = \frac{(n_1\lambda_1 + n_2\lambda_2)^2}{n_1\lambda_1^2 + n_2\lambda_2^2}. \tag{8}$$

The quantity c_{AM}^2 in (8) is the asymptotic-method approximation in [22], which is the exact theoretical value of c_∞^2 for the superposition process. The approximation is asymptotically correct for the mean delay in the queue as $\rho \to 1$ with n_1 and n_2 fixed (e.g., by increasing τ toward the critical value), by the heavy-traffic limit theorems [20]. The parameter v in (8) measures the effective number of component processes constituting the superposition process, e.g., if $n_2 = 0$, then $v = n_1$ and $w = [1 + 4(1 - \rho)^2(n_1 - 1)]^{-1}$. For the rest of this paragraph, suppose that $n_2 = 0$ (this is not strictly necessary). If $n_1 \to \infty$, then $w \to 0$ and $c_a^2 \to 1$, in accordance with Proposition 3). Even though the exact theoretical value $c_{1n_1}^2$ of the squared coefficient of variation of a single interval rapidly approaches 1 as n_1 increases, the QNA approximation selects an increasingly higher squared coefficient of variation c_a^2 as n_1 increases, to indirectly capture the effect of covariances (see Fig. 5).

Whenever it is known that the superposition of the data packet streams can be well approximated by a Poisson process, then $c_2^2 = 1$ and n_2 and λ_2 can be chosen arbitrarily so long as the product $n_2\lambda_2$ remains fixed. However, it is easy to see that the values of v, w, and c_a^2 in (8) are not independent of this choice. For this situation, the numerical results indicated that the approximation performs best if n_2 is selected so as to make $\lambda_2 \approx \lambda_1$. This is intuitively reasonable; we should expect the approximation to perform better if the processes being superposed are in the same time scale (i.e., have similar rates).

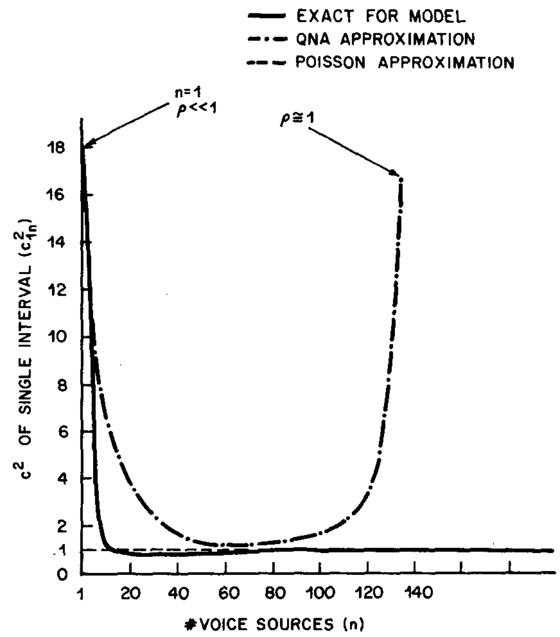


Fig. 5. Squared coefficient of variation of a single interval in the voice superposition process.

After we have applied (8), we can calculate congestion measures for the queue such as the mean and standard deviation of delay, the probability of delay, etc., by regarding it as a *GI/G/*1 queue (with renewal arrival process) partially characterized by λ, c_a^2, τ and c_s^2 where $c_a^2 \equiv c_a^2(\rho)$ is a function of ρ. See [28, Section 5.1] for specific formulas. See [29] and [40] for investigations of the quality of the approximations.

B. Numerical Comparisons

The parameters used in the numerical study are: mean talkspurt duration $\alpha^{-1} = 352$ ms, mean silence duration $\beta^{-1} = 650$ ms, and fixed packetization period $T = 16$ ms, so that the mean number of packets per talkspurt is 352/16 = 22 and the voice-line activity (the fraction of time each voice source is in talkspurt) is 0.351. Thus, for each voice line, the arrival rate is about 1320 packets per minute. The transmission line is assumed to be a T1 line with a rate of 1.536 Mbits/s. Thus, the capacity of the multiplexer without data is 136 voice lines (corresponding to nearly 100 percent line utilization).

1) The Simulation Experiments: Extensive simulations were performed to estimate the performance measures of this model. The simulation experiments were run on a Cray-1 computer with a special-purpose Fortran program.

To obtain the performance measures (expected delays, etc.), the multiplexer was simulated for fifteen minutes operating time. (The outputs printed out at one-minute intervals over the last six minutes indicated that the averages had stabilized.) Each experiment was repeated nine times with different seeds for the random number generator. For the case of 100 voice lines, the total number of arriving packets considered for each case was thus nearly 20 million (9 × 15 × 100 × 1320 = 17 820 000).

The estimates of the mean and standard deviation of packet delay were obtained by averaging the nine values associated with the different replications. The 95-percent confidence intervals were obtained assuming a Student-*t* distribution. A similar procedure was used when data traffic was also included.

To estimate the IDI (the curves in Fig. 4), the superposition arrival process for each n considered was simulated in one run of 300 min operating time. In the case of 60 voice lines, this means about 24 million packets. For larger values of k (e.g., $k > 100$), this amount of data is needed to produce reasonable estimates of c_k^2. Values of k up to 3000 were considered, but for the largest values of k the estimates had not yet stabilized. (The values where the estimates stabilized reasonably well are displayed in Fig. 4).

The estimate of c_k^2 in (1) as a function of k was calculated by collecting the arrival data over one-minute intervals and doing the calculations with this data for all k, $1 \leq k \leq 3000$. For each k, the interarrival times were grouped in adjacent nonoverlapping blocks of size k. (Nonoverlapping is not necessary.) For each k, the sums and sums of squares of these k-blocks (sums of k successive interarrival times) were then calculated for the data in the one-minute interval and accumulated. To be precise, let Y_{ki} be the length of the ith k block (i.e., $Y_{k1} = X_1 + \cdots + X_k$, $Y_{k2} = X_{k+1} + \cdots + X_{2k}$, etc.) and let n_k be the total number of k blocks. Our estimate of Var (S_k) is thus simply

$$n_k^{-1} \sum_{i=1}^{n_k} Y_{ki}^2 - \left(n_k^{-1} \sum_{i=1}^{n_k} Y_{ki}\right)^2. \tag{9}$$

We then combine this with the known exact value of $E(X_1)$ to obtain our estimate of c_k^2 in (1).

2) Evaluating the Approximations: The QNA analytic technique in Section III-A and [28] is compared to the Poisson approximation and simulation for the case of voice traffic only in Figs. 6 and 7 and Table III. (Delays refer to the time until beginning service; service times are not included. The packet service time is small compared to typical queueing delay values due to the high transmission rate of 1.536 Mbits/s.) The service times here are assumed to be deterministic. The 95-percent confidence intervals are indicated as well as the sample averages for the simulation. The QNA approach predicts the mean and standard deviation of delay in a packet-voice multiplexer with reasonable accuracy, especially at higher utilizations, where the Poisson approximation significantly underestimates both. However, the QNA approach does overestimate delays under moderate loads. From the analysis in Section II-B, we expected that the Poisson approximation would perform very well in light traffic and very poorly in heavy traffic, and that the QNA approximation would perform very well in heavy traffic. Figs. 6 and 7 confirm these predictions, and also indicate what happens over the entire range.

Fig. 8 shows the mean delay versus utilization curves for a multiplexer with voice and data. In this example, the voice load is fixed at 80 active voice lines (corresponding to a voice utilization of about 59 percent) and the data is varied. The data traffic is assumed to be characterized by Poisson arrivals and geometric packet size with mean 50 bytes. The Poisson arrival rate is determined by the data traffic intensity ρ_2 and the specified transmission rate of 1.536 Mbits/s. (The method outlined in Section III-A is valid, in general, for data traffic streams with renewal ar-

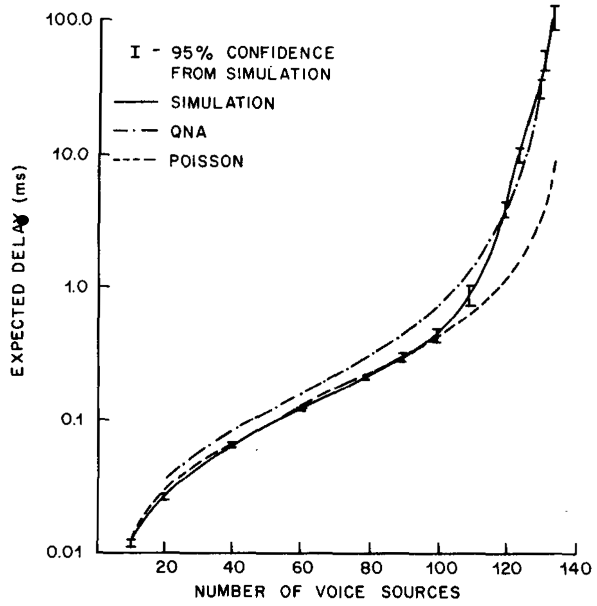


Fig. 6. Expected delay for a packet multiplexer (voice traffic only).

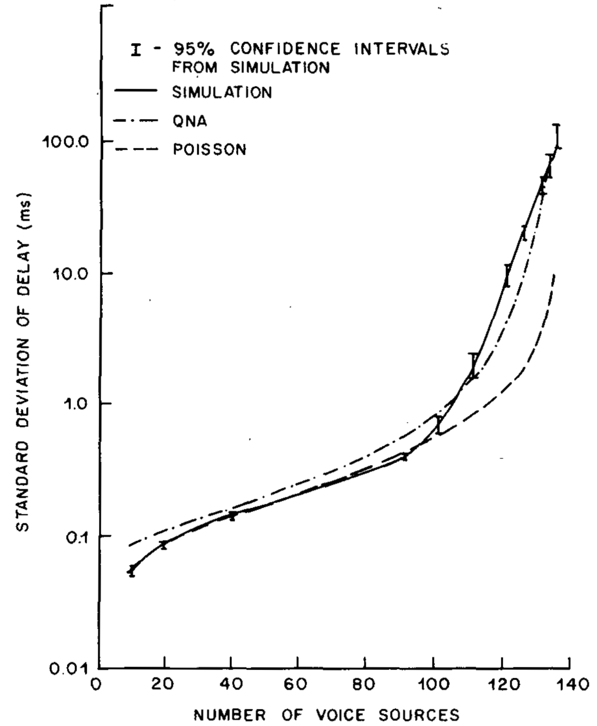


Fig. 7. Standard deviation of delay for a packet multiplexer (voice traffic only).

TABLE III
A Comparison of Approximations of the Mean Packet Delay to Simulation Estimates. (The Ratio of the Approximation to the Simulation Estimate is Given in Parentheses Under the Approximations. The 95-Percent Confidence Interval, Based on Nine Independent Simulation Replications, is Given in Parentheses Below the Simulation Estimates in Column 3.)

MODEL PARAMETERS		MODEL BEHAVIOR	APPROXIMATIONS					
			POISSON ARRIVAL	ORIGINAL QNA			MODIFIED QNA	
NUMBER OF VOICE SOURCES n	TRAFFIC INTENSITY ρ	SIMULATION (95% C.INT.)	M/D/1	η	$c_a^2(\rho)$	EW(ms)	$c_a^2(\rho)$	EW(ms)
20	0.146	0.03 (±0.001)	0.03 (1.00)	13.8	1.30	0.04 (1.33)	0.91	0.03 (1.00)
40	0.293	0.07 (±0.001)	0.07 (1.00)	19.5	1.22	0.08 (1.14)	0.95	0.07 (1.00)
60	0.439	0.13 (±0.001)	0.13 (1.00)	18.6	1.23	0.16 (1.23)	0.97	0.13 (1.00)
80	0.586	0.22 (±0.007)	0.24 (1.09)	13.6	1.31	0.31 (1.41)	0.98	0.23 (1.05)
90	0.659	0.31 (±0.02)	0.32 (1.03)	10.4	1.40	0.45 (1.45)	0.98	0.31 (1.00)
100	0.732	0.45 (±0.14)	0.46 (1.02)	7.1	1.58	0.72 (1.60)	0.98	0.45 (1.00)
110	0.805	0.89 (±0.14)	0.69 (0.78)	4.1	1.97	1.36 (1.53)	1.32	0.91 (1.02)
120	0.878	4.07 (±0.4)	1.20 (0.29)	1.76	3.12	3.75 (0.92)	3.44	4.14 (1.02)
125	0.915	10.4 (±1.3)	1.8 (0.17)	0.90	4.72	8.5 (0.82)	6.01	10.8 (1.04)
130	0.951	31.9 (±4.8)	3.3 (0.10)	0.30	8.71	28.4 (0.89)	10.98	35.9 (1.13)
132	0.966	52.1 (±7.5)	4.8 (0.09)	0.15	11.66	55.3 (1.06)	13.74	65.2 (1.25)
134	0.9807	109.6 (±21.4)	8.5 (0.08)	0.05	15.27	129.3 (1.18)	16.40	139.0 (1.27)

rival intervals.) The conclusions about the performance of the QNA technique are unchanged with the addition of data. However, it was observed (also see discussion in Section III-A) that the QNA approximation estimates mean delays best when $\lambda_1 \approx \lambda_2$. When data is Poisson it is always possible to pick n_2 and λ_2 so as to match λ_2 with λ_1 while keeping the overall data arrival rate $n_2\lambda_2$ fixed. The delays for voice and data are each approximated to be the delay seen by an arbitrary arrival. This is not always a good approximation, but it is reasonable for this model (see discussion in Section III-C-1) and Table IV).

C. *Refined Approximations*

The QNA methodology [28] can be used in two ways: First, it immediately provides relatively simple approximations for any model satisfying the basic assumptions; e.g., it can be applied directly, as described in [28] and Section III-A above. Second, it provides a starting point for developing better simple approximations, based on refinements obtained by exploiting special features of the particular problem. It should be expected that improved approximations can be obtained by making additional modifications, as we now illustrate for the packet-voice multiplexer.

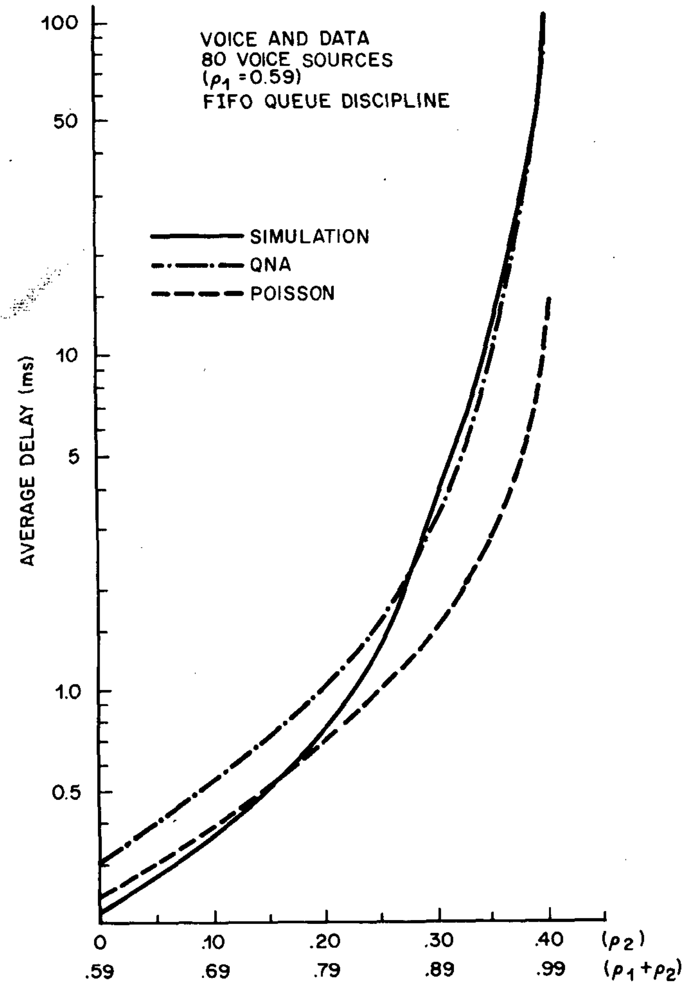


Fig. 8. Expected delay versus data utilization, ρ_2 (mixed voice and data traffic).

TABLE IV
Simulation Values of Expected Delays and Utilizations for the Example With Voice and Data

UTILIZATION			DELAYS			
				EXPECTED DELAYS		
DATA ONLY	TOTAL	TOTAL OBSERVED	PROBABILITY OF DELAY OBSERVED	VOICE ONLY	DATA ONLY	ARBITRARY PACKET
0.00	0.59	0.583	0.574	0.22	0.00	0.22
0.10	0.69	0.683	0.677	0.36	0.37	0.36
0.15	0.74	0.733	0.728	0.56	0.55	0.56
0.20	0.79	0.783	0.779	0.71	0.70	0.71
0.25	0.84	0.831	0.833	1.33	1.25	1.30
0.30	0.89	0.883	0.883	3.54	3.20	3.41
0.35	0.94	0.933	0.935	11.67	10.76	11.28
0.40	0.99	0.983	0.985	95.82	93.34	95.66

1) Probability of Delay: First, as a refinement to the probability of delay approximation in (48) of [28], we propose using the traffic intensity ρ instead. (This in turn also leads to a new approximation for the variance via [28, eq. (54)]. Since there are many component streams (e.g., 100), each stream tends to be a relatively negligible part of the whole, so that the delays experienced by packets are not greatly affected by other packets from the same source. The packet arrivals from a particular source tend to be like outside observers who do not influence the system. Hence, the proportion of arrivals that are delayed before beginning service should nearly coincide with the long-run proportion of time that the server is busy, which is exactly ρ. (A similar approximation for the probability of delay in a queue with a superposition arrival process was also developed empirically by Albin [29]. Based on many simulations of queues with superposition arrival processes, Albin suggested the pure stationary-interval method for approximating the probability of delay, which here yields $c_a^2 \approx 1.0$, as with Poisson arrivals. Then the probability of delay coincides with the fraction of time that the server is busy, ρ; see Wolff [41].)

For the same reason, in this problem the delays experienced by voice packets should be nearly the same as the delays experienced by arbitrary packets. Hence, we use the description of the delays experienced by an arbitrary packet also to describe the delays experienced by voice packets and data packets separately. Of course, in other situations these delays can be very different [42]. Table IV shows that the expected delays for voice and data tend to be very close, so that it is indeed appropriate to use the expected delay of an arbitrary packet to represent the expected delays of voice packets and data packets separately. Moreover, the probability of delay is very close to the traffic intensity. This illustrates how important insight into system behavior can be gained from the model structure, without obtaining numbers from simulations or system measurements.

2) Expected Delays: Comparisons to simulation in Fig. 6 show that, unlike the *M/G/*1 approximation, the QNA approximation accurately describes the dramatic increase in expected delays under heavy loads, but the QNA approximation seriously overestimates the expected delays under lighter loads. (From Table III we see that it is by as much as 60 percent.) Hence, now we propose a refined approximation, obtained by simply changing the weighting function $w(\rho, v)$ in (8), where we regard w as a function of a single variable $\eta = (1 - \rho)^2(v - 1)$. Heavy-traffic theory [31], [32] supports using η as a fundamental variable. In particular, η can be used to determine whether ρ is high relative to v. The key idea is that the *M/G/*1 approximation seems to be good for large η, say above 5, whereas the previous QNA approximation is pretty good for smaller values of η. In other words, the knee of the curve where the actual expected delay increases sharply, departing from the *M/G/*1 value, seems to occur for $\eta \approx 5$. (See Fig. 6 and Table III.) (Large-deviation theory, as in Weiss [12], should be helpful for locating this point more precisely, but we do not pursue this issue here.)

We want our new weighting function $\overline{w}(\eta)$ to have the following properties: 1) $\overline{w}(\eta) = 0$ for $\eta \geq 5$, 2) $\overline{w}(\eta)$ is a continuous decreasing function of η, 3) $\overline{w}(\eta) \to 1$ as $\eta \to 0$, and iv) $\overline{w}(\eta) \approx w(\eta)$ for $0 < \eta < 5$. The idea is to get something roughly appropriate. The particular weighting function we suggest is

$$\overline{w}(\eta) = \begin{cases} 0, & \eta \geq 5 \\ \dfrac{5 - \eta}{5 + 10\eta} = \dfrac{\tilde{w}(\eta) - \tilde{w}(5)}{1 - \tilde{w}(5)}, & 0 < \eta < 5, \end{cases} \tag{10}$$

where

$$\tilde{w}(\eta) = (1 + 2\eta)^{-1}. \tag{11}$$

We chose (10) because it satisfies requirements 1)–3) above. We chose the general form $(1 + A\eta)^{-1}$ in (11) to keep the same form as (8). We chose $A = 2$ to approximately equate $w(\eta)$ and $\overline{w}(\eta)$ for the special case $\eta = 2.5$, which falls in the middle of the interval (0, 5). Finally, we calculated other values of $\overline{w}(n)$ to check that it is reasonable.

Since the *M/G/*1 approximation slightly overestimates the expected delays under lighter loads, we also reduce the squared coefficient of variation c_a^2 when $\eta \geq 5$. In particular, instead of c_a^2, we suggest

$$\overline{c}_a^2 = \begin{cases} (n - 1)/(n + 1), & \eta \geq 5, n \geq 10, \\ 1 + \overline{w}(\eta)(c_{AM}^2 - 1), & 0 < \eta < 5. \end{cases} \tag{12}$$

We chose (12) for $\eta \geq 5$ because it seems consistent with the data and because it is a variant of the stationary-interval method [22]; see (6) and Table II. The case $0 < \eta < 5$ in (12) is obtained directly from (8), substituting $\overline{w}$ in (10) for w in (8).

The modified QNA approximation for expected packet delays exploiting (10)–(12) is compared with the original QNA approximation in Section III-A, the Poisson approximation and the simulation values in Table III. The improved accuracy of the new approximation is evident except at very high traffic intensities.

IV. A Time-Scale Test: The Related Blocking Model

We propose, as a general analysis technique, estimating the k-interval squared coefficient of variation (the IDI) in order to understand the nature of the dependence among successive interarrival times in a complex arrival process. Even the endpoints of the curve, c_1^2 and c_∞^2, can be very useful [22]. (In many cases, as here, these can be determined analytically.) We contend that curves such as are displayed in Fig. 4 can provide important insight in many contexts. For example, here Fig. 4 suggests that the superposition of 60 voice streams behaves like a Poisson process in a sufficiently short time scale (e.g., less than 20 interarrival times), but like a highly variable non-Poisson process in a longer time scale (e.g., of more than 100 interarrival times). This interpretation is supported by the delay curves in the numerical examples of Section III.

In this section, we describe another test. We consider a variant of the same model with the same superposition packet arrival process, modified by imposing a limit to the number of waiting spaces in the queue, i.e., by having a finite buffer. (We assume that, when the buffer is full,

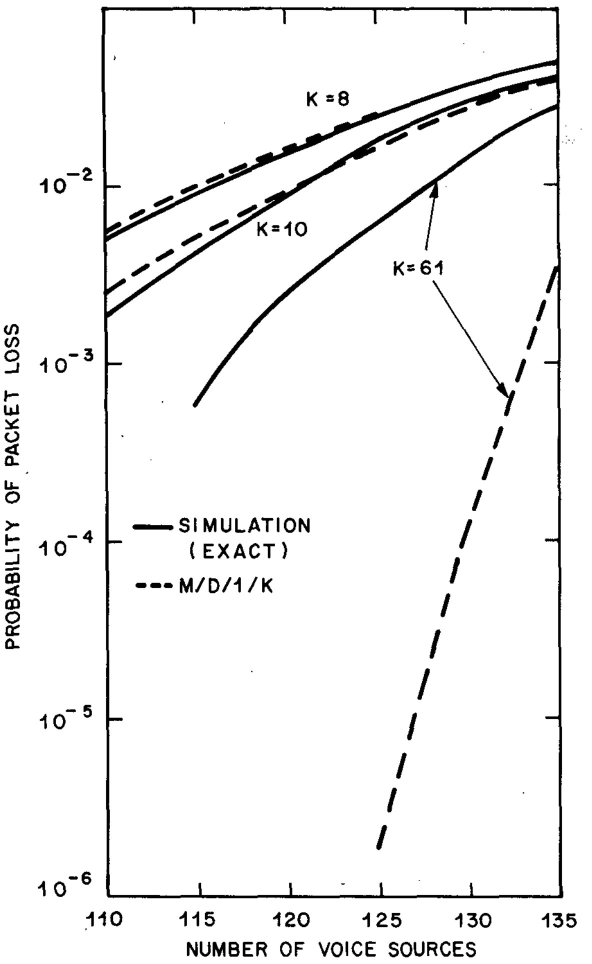


Fig. 9. Voice packet loss—Comparison of exact and *M/D/1/K* models for different buffer sizes, *K*.

arrivals are lost without generating retrials). Fig. 4 leads us to expect that the actual proportion of packets lost would be nearly the same as for a Poisson arrival process (the *M/D/1/K* model) when the buffer-size *K* is sufficiently small, but much greater than predicted by the Poisson model when *K* is large. We would expect this because a buffer of size *K* means that at most *K* consecutive interarrival times can interact directly in the queue. When *K* is large and the offered load is high, the queue lengths are often large and hence many interarrivals times interact in the queue. Then the burstiness of the arrival process due to long-term positive covariances (see Fig. 4 and Section II-B) should cause a much higher packet loss compared to a Poisson model. On the other hand, when *K* is fairly small, few interarrival times can interact in the queue. Then the effect of positive covariances should be negligible and the exponential nature of the single interval (see Appendix A and flat portion of curve in Fig. 4) should dominate. As a result the packet loss in the system should be comparable to that of an *M/D/1/K* model.

We tested this hypothesis by performing simulation experiments with different buffer sizes. The actual proportions of packets lost are compared with the proportions lost when there is a Poisson arrival process in Fig. 9. The results are obviously consistent with our expectations. For a buffer size of 61, the actual proportion of packets lost is much greater than would be the case with a Poisson arrival process, but for buffer sizes of 8 and 10, the proportion of packets lost nearly coincides with what would be the case with a Poisson arrival process. In fact, for buffer sizes of 8 and 10, the Poisson blocking probability is actually slightly *greater* under lighter loads. This result is not surprising, as Fig. 4 and Table II indicate that a single interarrival-time distribution may be even less variable than an exponential. Thus when *K* is small, the packet loss may indeed be even smaller than that of the *M/D/1/K* model. (In actual systems, the number of buffers typically would be large (about 60), so that burstiness due to long term positive covariances must be taken into consideration for delay as well as packet loss.) We do not yet know how to determine exactly when the Poisson approximation ceases to be appropriate, but the IDI clearly provides important insight.

V. Conclusions

The specific focus of this study has been on the performance of a statistical multiplexer for voice and data. We have obtained useful approximations and important insights for this model, but there also are important implications for the performance analysis of other complex queueing systems. There are three themes in this paper: 1) using relatively simple approximations, 2) analyzing arrival processes to better understand the nature of the dependence among successive interarrival times, 3) identifying the relevant time scale.

The first theme is relatively well understood: Simple approximations are obviously useful when they are sufficiently accurate; we have shown that simple approximations with sufficient accuracy can indeed be developed for this model. The second theme about dependence among successive interarrival times has a long history in voice traffic theory, most commonly involving the concept of peakedness; see Eckberg [43] and references cited there. However, dependence among the interarrival times in arrival processes has not yet received as much attention as it should in the performance analysis of packet networks and computer systems. Of course, many system measurements have been made [44]–[47] and many simulation models have been built to obtain data about system performance, but only rarely are the arrival processes measured (or modeled and studied analytically) with the intent of understanding the statistical fluctuations due to dependence among interarrival times. This paper clearly demonstrates that dependence among interarrival times can play an essential role. Our analysis of the multiplexer model shows that the aggregate packet arrival process possesses exceptional long-term positive dependence, partially characterized by the indexes of dispersion and that this dependence is a major cause of congestion in the multiplexer queue under heavy loads. Even though the aggregate packet arrival process with many components is nearly Poisson in the sense of Proposition 3, it is not appropriate to simply model the aggregate packet arrival process as a Poisson process with the correct rate, cf. [47]; the $M/G/1$ model does not work under heavy loads.

We recommend routinely measuring the indexes of dispersion and plotting curves such as in Figs. 3 and 4. A statistical estimation procedure is easily implemented in both system measurements and computer simulations (when ample data are available). For complex models we recommend estimating the indexes of dispersion of several flows in the model in order to help reveal how the dependence in an initial arrival process is altered by system features. For example, in systems with windows and other flow-control mechanisms we can see how these features alter the dependence structure; i.e., reduce variability. We can also identify sources of unanticipated and undesired variability, so that it can be eliminated or appropriately controlled. We believe that the indexes of dispersion should become standard measurement tools in performance analysis.

Finally, there is the third theme: identifying the relevant time scale. The third theme comes to the fore when we try to relate the IDI to the congestion in the queue. By the *relevant time scale* here we mean the typical durations over which arrivals interact in a queue, and hence collectively influence the queue behavior. We see that an appropriate approximation for the arrival process depends on the time scale, and the relevant time scale in turn depends on the traffic intensity in the queue [31], [32], [34]. The concept of the relevant time scale was well illustrated with the blocking model in Section IV.

A direction for future research is to develop approximations for congestion measures in a queue when the arrival process is partially characterized by its average arrival rate λ and the IDI $\{c_k^2, k \geq 1\}$; e.g., $c_a^2 \equiv c_a^2(\rho) = \Sigma_{k=1}^{\infty} c_k^2 w(k, \rho)$, where $w(k, \rho)$ is a more general weighting function than in (8) and $\tau = 1$. The idea is to develop a formula that depends on λ and $\{c_k^2, k \geq 1\}$, but *not* directly on other parameters specific to the structure of the system.

Appendix

The Interarrival-Time Distribution in the Superposition Process

In Section II-B, we indicated that the aggregate packet arrival process resulting from the superposition of voice sources differs from a Poisson process primarily because of the dependence among successive intervals. We substantiate this claim now by calculating the cdf $F_n(t)$ and the squared coefficient of variation c_{1n}^2 for a single interarrival time in the aggregate packet arrival process as a function of the number n of voice sources. For the values of n we are considering, the distribution is nearly exponential and the squared coefficient of variation is nearly one. The numerical values are shown in Fig. 5 and Tables I and II. In fact, for sufficiently large n, the interarrival-time distribution is actually somewhat *less* variable than an exponential distribution with the same mean; i.e., it has less mass near zero and at large values. This property is reflected by the squared coefficient of variation c_{1n}^2, which is actually less than one for $n \geq 12$. It decreases from 18.1 for $n = 1$ to a minimum of 0.91 at $n = 18$ and then increases toward 1 as $n \to \infty$.

Let $F(t)$ be the interarrival-time cdf in the renewal process for one voice source with mean λ^{-1} and let $F_e(t)$ be the associated cdf of the equilibrium-excess variable, defined as usual by

$$F_e(t) = \lambda \int_0^t [1 - F(u)]\, du, \qquad t \geq 0, \tag{A.1}$$

see [22, eq. (1.5)]. It is well known that the cdf of a single interarrival time in the superposition process is

$$F_n(t) = 1 - (1 - F(t))(1 - F_e(t))^{n-1}, \qquad t \geq 0, \tag{A.2}$$

which is a special case of (4.4) in [22].

Here, the interarrival-time cdf from one source has tail

$$1 - F(t) = \begin{cases} 1 & 0 \leq t \leq T \\ (1 - p)\, e^{-\beta(t-T)}, & t \geq T \end{cases} \tag{A.3}$$

and mean

$$\lambda^{-1} = T + \frac{1-p}{\beta} = \frac{\beta T + 1 - p}{\beta}, \tag{A.4}$$

so that

$$1 - F_e(t) = \begin{cases} 1 - \lambda t, & 0 \leq t \leq T, \\ \dfrac{1-p}{\beta}\, e^{-\beta(t-T)}, & t \geq T. \end{cases} \tag{A.5}$$

Hence,

$$1 - F_n(t) = \begin{cases} (1 - \lambda t)^{n-1}, & 0 \leq t \leq T \\ \dfrac{(1-p)^n\, e^{-\beta n(t-T)}}{(\beta T + 1 - p)^{n-1}}, & t \geq T. \end{cases} \tag{A.6}$$

If we appropriately scale time by n in the superposition process, then

$$1 - F_n(t/n) = \left(1 - \frac{\lambda t}{n}\right)^{n-1}, \qquad 0 \leq t \leq nT, \tag{A.7}$$

so that

$$1 - F_n(t/n) \to e^{-\lambda t} \quad \text{as} \quad n \to \infty \quad \text{for any} \quad t > 0. \tag{A.8}$$

Hence, for large n, only the first component in (A.6) is relevant, and it is the geometric approximation of the exponential tail; [48, p. 1].

In Table I, we compare the tail probabilities in (A.6) to an exponential distribution with the same mean for our multiplexer model in the cases $n = 20$ and 100. The close fit is apparent. Also note that the exponential distribution has more mass near zero and at large values, as previously claimed.

From (A.6), it is straightforward to calculate the squared coefficient of variation. Recall that the mean of an interarrival time X_n in the superposition process is

$$EX_n = \frac{EX_1}{n} = \left(\frac{\beta T + 1 - p}{n\beta}\right). \qquad \text{(A.9)}$$

Using the formula for the cdf $F_n(t)$ of X_n in (A.6), we obtain

$$\begin{aligned} E(X_n^2) &= 2\int_0^\infty t[1 - F_n(t)]\,dt \\ &= 2\int_0^T t\left(1 - \frac{\beta t}{\beta T + 1 - p}\right)^{n-1} dt \\ &\quad + \frac{2(1-p)^n e^{\beta nT}}{(\beta T + 1 - p)^{n-1}} \int_T^\infty te^{-\beta nt}\,dt. \qquad \text{(A.10)} \end{aligned}$$

In turn,

$$\begin{aligned} &\int_0^T t\left(1 - \frac{\beta t}{\beta T + 1 - p}\right)^{n-1} dt \\ &= \left(1 - \frac{\beta T}{\beta T + 1 - p}\right)^{n+1} \left(\frac{1}{(n+1)\left(\frac{\beta}{\beta T + 1 - p}\right)^2}\right) \\ &\quad - \left(1 - \frac{\beta T}{\beta T + 1 - p}\right)^{n} \frac{1}{n\left(\frac{\beta}{\beta T + 1 - p}\right)^2} - \frac{1}{(n+1)\left(\frac{\beta}{\beta T + 1 - p}\right)^2} \frac{1}{n\left(\frac{\beta}{\beta T + 1 - p}\right)^2} \\ &= \left(\frac{\beta T + 1 - p}{\beta}\right)^2 \left[\frac{2n}{n+1} + \left(\frac{1-p}{\beta T + 1 - p}\right)^{n-1} \frac{1}{n+1} - \left(\frac{1-p}{\beta T + 1 - p}\right)^{n} \frac{1}{n}\right] \end{aligned}$$

and $\int_T^\infty te^{-\beta nt}\,dt = [1/(n\beta)^2]\,e^{-\beta nT}(1 + \beta nT)$, so that, after some algebra, we obtain

$$c_{1n}^2 = 1 - \frac{2}{n+1} + \left(\frac{1-p}{\beta T + 1 - p}\right)^{n+1} \left(\frac{2}{1-p} - \frac{2n}{n+1}\right). \qquad \text{(A.11)}$$

In our application, $p = 21/22$, $T = 16$ and $\beta = 1/650$, so that $(1 - p)/(\beta T + 1 - p) = 0.6489$ and

$$\begin{aligned} c_{1n}^2 &= 1 - \frac{2}{n+1} + (0.64898)^{n+1}\left(44 - \frac{2n}{n+1}\right) \\ &\approx 1 - \frac{2}{n+1} \quad \text{for large } n. \qquad \text{(A.12)} \end{aligned}$$

Table II and Fig. 5 show c_{1n}^2 as a function of n.

Acknowledgment

We thank D. M. Lucantoni and H. Heffes for helpful discussions and for providing the index of dispersion for counts shown in Fig. 3. We also thank A. Anastasio for providing skillful assistance with the computer simulations. Finally, we thank the referees for many helpful suggestions.

References

[1] S. J. Campanella, "Digital speech interpolation," *COMSAT Tech. Rev.*, vol. 6, pp. 127-158, Spring 1976.

[2] S. R. Amstutz, "Burst switching—A method for distributed and integrated voice and data switching," *IEEE Commun. Mag.*, pp. 36-42, Nov. 1983.

[3] K. Sriram, P. K. Varshney, and J. G. Shanthikumar, "Discrete-time analysis of integrated voice-data multiplexers with and without speech activity detectors," *IEEE J. Select. Areas Commun.*, vol. SAC-1, no. 6, pp. 1124-1132, Dec. 1983.

[4] P. O'Reilly, "A fluid-flow approach to performance analysis of integrated voice-data systems with DSI," in *Proc. Int. Conf. Modeling and Tools for Performance Anal.*, Sophia-Antipolis, France, June 1985, pp. 121-136.

[5] J. S. Turner and L. F. Wyatt, "A packet network architecture for integrated services," in *Proc. GLOBECOM 83*, San Diego, CA, Nov. 1983, pp. 2.1.1-6.

[6] W. Hoberecht, "A layered network protocol for packet voice and data integration," *IEEE J. Selected Areas in Commun.*, vol. SAC-1, no. 6, pp. 1006-1013, Dec. 1983.

[7] J. J. Kulzer and W. A. Montgomery, "Statistical switching architectures for future services," in *Proc. ISS '84*, Florence, Italy, May 1984, pp. 43A.1.1-6.

[8] J. S. Turner, "Design of an integrated services packet network," presented at the Ninth Data Commun. Symp., Whistler Mt, B.C., Canada, Sept. 1985.

[9] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, pp. 1871-1894, Oct. 1982.

[10] J. Sequel, Y. Tanaka, and M. Akiyama, "Simulation analysis of the waiting time distribution of a packetized voice concentrator," *Trans. IECE Japan*, vol. E 68, pp. 115-122, Feb. 1982.

[11] Y. K. Tham and J. N. P. Hume, "Analysis of voice and low-priority data traffic by means of brisk periods and slack periods," *Comput. Commun.*, vol. 6, no. 1, pp. 14-22, Feb. 1983.

[12] A. Weiss, "A new technique for analyzing large traffic systems," *J. Appl. Prob.*, June 1986.

[13] T. E. Stern, "A queueing analysis of packet voice," in *Proc. IEEE Global Telecomm. Conf.*, San Diego, CA, Dec. 1983, pp. 2.5.1-2.5.6.

[14] T. Suda, H. Miyahara, and T. Hasegawa, "Performance evaluation of a packetized voice system—Simulation study," *IEEE Trans. Commun.*, vol. COM-32, pp. 97-102, May 1984.

[15] Y. C. Jenq, "Approximations for packetized voice traffic in statistical multiplexer," *Proc. IEEE INFOCOM*, Apr. 1984, pp. 256-259.

[16] M. L. Luhanga, "Analytical model of a packet voice concentrator," Columbia University, New York, NY, Tech. Rep. No. 1984-02.

[17] S. Ganguly, "Analysis and performance evaluation of multiplexed packet voice and data," Ph.D. dissertation, Dep. Elec. Eng., Columbia University, New York, NY, June 1985.

[18] J. N. Daigle and J. D. Langford, "Queueing analysis of a packet voice communication system," *Proc. IEEE INFOCOM*, Washington, DC, Mar. 1985.

[19] D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*. London, England: Methuen, 1966.

[20] D. L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, II: Sequences, networks and batches," *Adv. Appl. Prob.*, vol. 2, no. 2, pp. 355-369, Autumn 1970.

[21] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[22] W. Whitt, "Approximating a point process by a renewal process: Two basic methods," *Oper. Res.*, vol. 30, no. 1, pp. 125-147, January-February 1982.

[23] H. Heffes, "A class of data traffic processes—Covariance function characterization and related queueing results," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 897-929, July-Aug. 1980.

[24] H. Heffes and D. M. Lucantoni, "A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," this issue, pp. xxx-xxx.

[25] V. Ramaswami, "The *N/G/*1 queue and its detailed analysis," *Adv. Appl. Prob.*, vol. 12, no. 1, pp. 222-261, Mar. 1980.

[26] W. Whitt, "Weak convergence theorems for priority queues: Preemptive resume discipline," *J. Appl. Prob.*, vol. 8, no. 1, pp. 74-94, Mar. 1971.

[27] J. M. Harrison, "A limit theorem for priority queues in heavy traffic," *J. Appl. Prob.*, vol. 10, no. 3, pp. 613-629, Sept. 1973.

[28] W. Whitt, "The queueing network analyzer," *Bell Syst. Tech. J.*, Part 1, vol. 62, no. 9, pp. 2779-2815, Nov. 1983.

[29] S. L. Albin, "Approximating a point process by a renewal process, II: Superposition arrival processes to queues," *Oper. Res.*, vol. 32, pp. 1133-1162, Sept.-Oct. 1984.

[30] S. L. Albin, "On Poisson approximations for superposition arrival processes in queues," *Management Sci.*, vol. 28, no. 2, pp. 126-137, Feb. 1982.

[31] G. F. Newell, "Approximations for superposition arrival processes in queues," *Management Sci.*, vol. 30, pp. 623-632, May 1984.

[32] W. Whitt, "Queues with superposition arrival processes in heavy traffic," *Stoch. Proc. Appl.*, vol. 21, pp. 81-91, 1985.

[33] E. Cinlar, "Superposition of point processes," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. W. Lewis, Ed. New York: Wiley, 1972, pp. 549-606.

[34] W. Whitt, "Departures from a queue with many busy servers," *Math. Oper. Res.*, vol. 9, no. 4, pp. 534-544, Nov. 1984.

[35] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell Syst. Tech. J.*, vol. 47, no. 1, pp. 73-91, Jan. 1968.

[36] J. G. Gruber, "A comparison of measured and calculated speech temporal parameters relevant to speech activity detection," *IEEE Trans. Commun.*, pp. 728-738, Apr. 1982.

[37] Y. Yatsuzuka, "Highly sensitive speech detector and high-speed voiceband discriminator in DSI-ADPCM system," *IEEE Trans. Commun.*, vol. COM-30, pp. 739-750, Apr. 1982.

[38] C. J. May and T. J. Zebo, private work, AT&T Bell Lab., 1981.

[39] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*. New York: Academic, 1975.

[40] J. G. Klincewicz and W. Whitt, "On approximations for queues, II: Shape constraints," *AT&T Bell Lab. Tech. J.*, vol. 63, no. 1, pp. 139-161, Jan. 1984.

[41] R. W. Wolff, "Poisson arrivals see time averages," *Oper. Res.*, vol. 30, pp. 223-231, Mar.-Apr. 1982.

[42] A. Kuczura, "Queues with mixed renewal and Poisson inputs," *Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1305-1326, July-Aug. 1972.

[43] A. E. Eckberg, "Generalized peakedness of teletraffic processes," in *Proc. Tenth Int. Teletraffic Cong.*, Montreal, P.Q., Canada, June 1983, p. 4.4b.3.

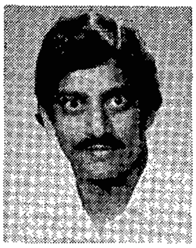
[44] E. Fuchs and P. E. Jackson, "Estimates of distributions of random variables for certain communications traffic models," *Commun. ACM*, vol. 13, pp. 752-757, 1970.

[45] P. F. Pawlita, "Traffic measurements in data networks, recent measurement results, and some implications," *IEEE Trans. Commun.*, vol. 29, pp. 525-535, 1981.

[46] S. P. Morgan, private work, AT&T Bell Lab., 1982.

[47] B. G. Kim, "Characterization of arrival statistics of multiplexed voice packets," *IEEE J. Select. Areas Commun.*, vol. SAC-1, no. 6, pp. 1133-1139, Dec. 1983.

[48] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Second Ed. New York: Wiley, 1971.



Kotikalapudi Sriram (S'80-M'83) received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1977 and 1979, and the Ph.D. degree in electrical engineering from Syracuse University, Syracuse, New York, in 1983. He was awarded the Syracuse University Fellowship during 1981-1982.

He was a Visiting Assistant Professor in the Department of Electrical and Computing Engineering at Syracuse University in 1982-83. He has been with AT&T Bell Laboratories, Holmdel, New Jersey, since October 1983. He is currently a Member of Technical Staff in the Department of Teletraffic Theory and System Performance. His research interests are in communication theory, communication networks, and speech processing.

Ward Whitt received the A.B. degree in mathematics from Dartmouth College, Hanover, NH, in 1964 and the Ph.D. degree in operations research from Cornell University, Ithaca, NY, in 1969.

He taught in the Department of Operations Research at Stanford University in 1968-1969, and in the Department of Administrative Sciences at Yale University from 1969-1977. Since 1977, he has been employed by AT&T Bell Laboratories, Holmdel, NJ. He is currently a Member of Technical Staff in the Department of Operations Research, Systems Analysis Center, where the primary mission is to investigate and improve the AT&T product realization process.

Dr. Whitt is a member of the Operations Research Society of America and the Institute of Mathematical Statistics.